

接続詞と主辞に着目した特許文の並列構造解析

Parallel structure analysis of patent sentences concerning conjunctions and head words

山形大学工学部技術部情報技術室技術員 **高橋 尚矢**

PROFILE

1989年生。2012年山形大学工学部卒。同年山形大学工学部技術部情報技術室に採用。

✉ naoya_takahashi@yz.yamagata-u.ac.jp TEL 0238-26-3339

山形大学大学院理工学研究科教授 **横山 晶一**

PROFILE

1949年生。1972年東京大学工学部卒。同年電子技術総合研究所入所。1991年同所知能情報部自然言語研究室長。1993年4月より山形大学。現在大学院理工学研究科（情報科学分野）教授。工学博士。アジア太平洋機械翻訳協会（AAMT）Japio 特許翻訳研究会副委員長。

✉ yokoyama@yz.yamagata-u.ac.jp TEL 0238-26-3336

1 はじめに

近年、国際的な特許の共有化に伴い国際特許の申請数も増加し続けている。特許文の検索や翻訳などの作業には多くの人手が必要であり、そのため作業を自動化または半自動化することが求められている。これらを解決するためには特許文中に含まれる情報を的確に抽出することが要求される。したがって、特許文に対する正確な係り受け解析が不可欠である。

特許文の課題や解決手段の部分は、200文字を超える長大な一文になることが多い。しかも単語同士の係り受けが複雑であり、意味が明確でないことがある。これは特許文独特の記述や専門用語の多さなどが原因である。そのため通常文に比べると係り受けが曖昧になりやすく、解析の誤りが発生しやすいという特徴がある。

本稿では、特許文中にあらわれる助詞の「に」に着目する。助詞の「に」には格助詞的用法と並列助詞（並立助詞という表現の方が正しいが、今回は並列構造を捉えるので並列助詞と記す^[1]）的用法があるが、特許文において並列的な用いられ方をすることは少ないと思われる。しかしKNP^[2]では「に」を並列的用法だと誤解析する例が見受けられたため、原因の解明と修正する方法を検討する。

また並列助詞である「と」と主辞との関連から大きな

修飾構造を解明し、並列修飾句をよりの確に把握できるシステムを作成する予定である。

2 関連研究

松山ら^[3]は特許文ではなく法令文書を対象にした並列構造解析をおこなっている。あらかじめ法令文書の重要語を決め、それらをもとに並列構造を検出した。この研究は並列構造を pfi（前方句、iは前方句の数だけ増える）、key（並列となるキーワード）、pb（後方句）と定義し、この型に当てはまる部分を検出するというものである。具体例を挙げると、

“四百八十から保険料納付済期間の月数（pf3）、保険料四分の一免除期間の月数（pf2）、保険料半額免除期間の月数（pf1）及び（key）保険料四分の三免除期間の月数（pb）を合算した月数を控除して得た月数を限度とする”

というふうに法令文では3つ以上の句が並列関係にある場合は2つ以上の前方句を読点で並べた後、並列キー、後方句が続くという特徴がある。

しかし階層的並列構造は検出が難しい、前方句と後方句の長さが違いすぎると並列構造の検出に失敗しがちで

ある、など改良の余地がある。また法令文書では並列を意味する語が細かい規則で設定されている^[4]のに対し、特許文ではそのような規則が無いため並列階層の特定が困難である。

我々は、これまでに特許文の並列構造解析について様々な手法で解析を試みてきた^{[5][6]}。ただし主辞に限定した解析のため、特許文全体としては精度がさほど上がらなかった。今回は主辞をベースにしつつも助詞（格助詞・並列助詞）の観点から精度向上を目指す。

3 助詞「に」の用法および解析例

助詞・助動詞の辞典^[7]によれば、並列助詞の「に」

とは「AにB」の形で両者を対比したり両者の取り合わせを問題にするようなセットとなることを表す」とある。具体例としては弱り目に祟り目、お弁当にお茶、などペアであることが重要視される文節間にあらわれる。これらの表現は文語的で特許文（説明文）中に登場するとは考えにくい。

図1に示したのは助詞の「に」がKNPで誤って出力した例である。元の文は、

“培土板の上部に、培土板の前方に向けてほぼ水平に延び、下方からの迫り上げ土圧により後部の取付部を支点として弾性をもって変位する可動邪魔板を設ける。”

とあり、『培土板の上部に』が主部に、『培土板の前方に

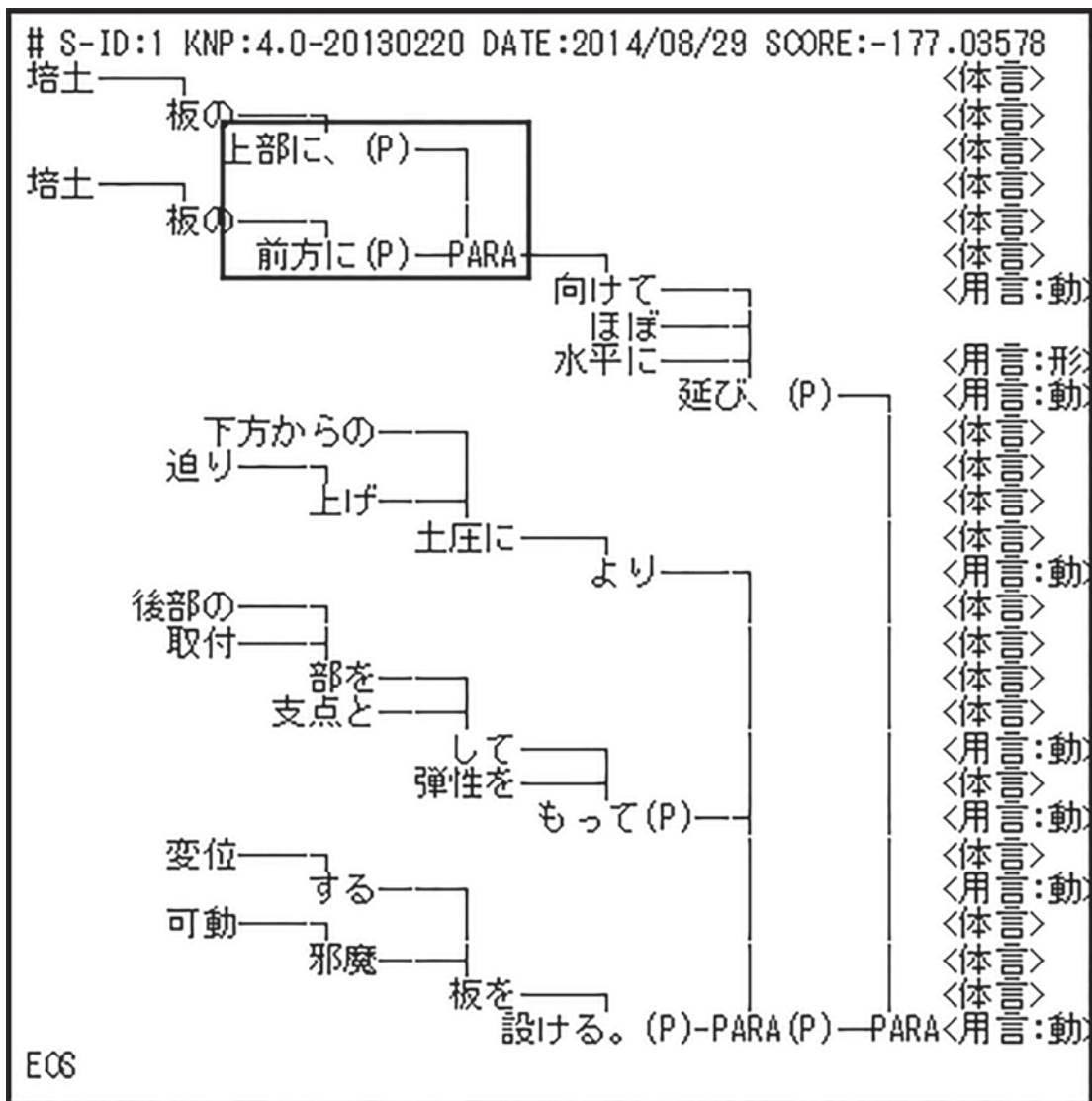


図1 助詞「に」を誤解析した例

向けてほぼ水平に延び、下方からの迫り上げ土圧により後部の取付部を支点として弾性をもって変位する可動邪魔板を設ける』が述部になるはずである。しかし図1の太線で囲った部分を見ると、『培土板の上部』と『培土板の前方』が並列構造を持つ、と誤解析していることがわかる。この誤解析の原因は『上部に』の「に」を並列助詞として解釈していると考えられる。

図2に示したのは意味自体は問題ないが解析の観点からは不都合がある例である。この文は、

この発明は、携帯機器等や、電子手帳または、各種表示装置に使用されているタッチパネルに関し、とくにその超音波反射波防止構造に関する。

のように「に」の部分¹が修飾句的²使われ方をしている。

図2の太線で囲まれた部分は正しくは、

[{携帯機器等} や {電子手帳} または {各種表示装置}] に使用されている・・・

という意味合いになる。この箇所は意味は合っているのだが、助詞の「に」が格助詞ではなく並列助詞扱いとなっている（PARAでつながれた部分）。これはKNPの仕様として文節ごとに区切りを設けるためであり、このような例も意味解釈さえ合っていれば正解扱いとした。

今回はランダムに選択した特許文の中から助詞の「に」を含む文章85文を人手で検証したところ、助詞の「に」が原因で誤解析した例が13件見受けられた。うち一つの文中で助詞の「に」が1回しか現れない文50文における誤解析は4文だったのに対し、一つの文中に助

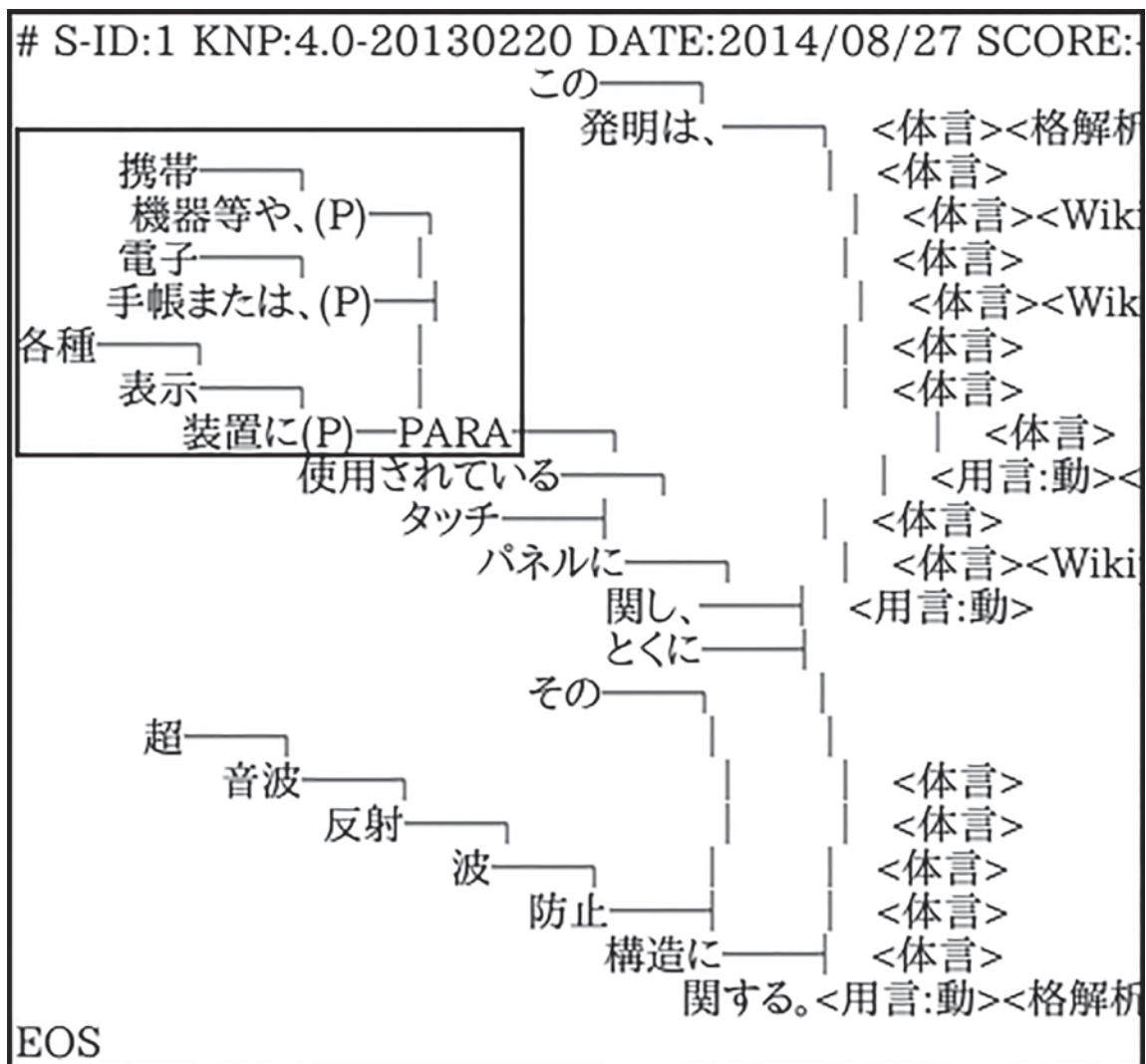


図2 意味合いは合っているが解析上問題がある例

詞「に」が2回以上現れる文35文における誤解析は9文と1回のみ登場する場合に比べ明らかに高かった。

4 今後の方針

今回の調査で、構造解析システムは長文の解析が不完全であることが改めてわかった。今後はKNPで特許文を解析する際に助詞「に」を正しく解析できるような手法を引き続き検討し、それに加え並列助詞「と」と主辞の関係性を探る。KNPで並列修飾句を的確に扱えるような手法を検討し、特許文における修飾構造を正しく解析するシステムを作成する予定である。

具体的には、あらかじめ誤解析を起こす確率の高い文章が発見できれば未然に誤解析を防ぐ手立てができると考え、特許文のうち1文中に助詞の「に」が2回以上登場する文を収集し、誤解析を示す文の法則性を求める。まずは教師あり学習でどの程度間違いを修正できるか実験する予定である。

参考文献

- [1] 日本語文法学会 編：日本語文法辞典、大修館書店 (2014)
- [2] 日本語構文・格解析システム KNP：<http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP>
- [3] 松山宏樹、白井清昭、島津明：法令文書を対象にした並列構造解析、言語処理学会 第18回年次大会 (2012)
- [4] 岩本秀明、長野馨、永井秀利、中村貞吾、野村浩郷：法律文における並列構造の特徴とそれに基づく制限言語モデルについて、情報処理学会自然言語処理研究会 (1993)
- [5] 横山晶一：接尾辞に着目した特許文の並列構造解析、Japio YEAR BOOK 2012 pp.250-253
- [6] 高橋尚矢、横山晶一：特許文における入れ子構造の調査、Japio YEAR BOOK 2013 pp.266-270
- [7] 森田良行 著：助詞・助動詞の辞典、東京堂出版 (2007)